Sheila M. Reynolds
Lecture 16, 11/22/04
CSE 527

Continuing discussion of "Pre-mRNA Secondary Structure Prediction Aids Splice Site Recognition" (a presentation to the Pacific Symposium on Biocomputing, Jan 2002, by D.J.Patterson, K.Yasuhara, and W.L.Ruzzo).

- The goal is to add information about the predicted secondary structure of the mRNA to the task of recognizing splice sites. The structure information is in *addition* to the primary sequence information.

- Training data set consisted of *equal* numbers of known splice sites and non-sites selected randomly with the following criteria: aligned to "AG" consensus, and located within 100 bases of an annotated splice site. (All training segments were 100 bases long, with the candidate splice-site in the middle.)

- Note that the fact that the training data consisted of equal numbers of splice-sites and non-sites is a potential problem as many more non-sites would be expected in "real" data.

- "MFOLD" was used for secondary structure prediction. The structure prediction needed to be summarized in some manner. The following three features were extracted:

  1. optimal folding energy (the energy of the most stable predicted structure)
  2. "max helix score" -- indicating how likely each position is to be near (within [-5,+5] positions) a helix structure. (This results in a vector of 100 values.)
  3. neighbor-pairing-correlation-model (NPCM) :
     - a new sequence was created based on the predicted structure:
       - "O" : unpaired base
       - "P" : paired base
       - "S" : paired *and* stacked base
     - using these modified training sequences, a pair of 2nd order Markov Models were built (one model trained on known splice-sites, the other model trained on the "non-sites") -- each test-sequence was then scored based on these models using the log-likelihood ratio.

- These statistics (the above 3 along with the sequence-based metric discussed in the previous lecture) were then fed into two types of machine-learning algorithms: a Decision Tree (Quinlan's C4.5) and a Support-Vector-Machine (Noble's SVM 1.1 with Radial Basis Kernel degree 2). Both of these algorithms take a vector of statistics and produce a yes/no answer.

- Testing / Training was done using 10-fold cross-validation (90% of the data is used for training, and 10% for testing, and this is repeated 10 times so that all of the data has been used for testing). The Wilcoxon test (which assumes that the improvement statistics are symmetric but not necessarily Gaussian) was used to compute a p-value to indicate whether the improvement seen (as compared to splice-site recognition based on primary sequence information alone) was consistent across all 10 subsets.

- Results were shown using Decision Trees. The baseline mean accuracy was 92.73% (using only the primary sequence information and the Weight-Array-Matrix). The best score obtained using the Optimal-Free-Energy score and the Max-Helix score in addition to the primary sequence WAM was 93.21% (providing a reduction in the error rate of 6.6%).

- Looking at some of the results related to structure prediction:
  - 25% more likely for splice sites to pair at position -2
  - 35% more likely for splice sites to initiate a helix at position -2 and -1
  - 45% more likely for splice sites to continue a helix through the splice site

- Splice sites were nearly twice as likely to be in a helix than not (63% to 37%) whereas the non-sites used in training were equally likely to be in a helix or not (52% to 48%). Within the splice-site helices there was also a slight tendency to fold left rather than right. This helix result was somewhat counter-intuitive. One might have thought that the helix structure might have prevented the pre-mRNA from hybridizing with the small nuclear RNAs that are part of the splicing machinery, but the presence of a short helix may actually help with the hybridization process.

- These structure-related statistics are adding information about long-distance relationships into the splice-site model than a more conventional nth order model (n small) would probably not be able to describe.

Support-Vector-Machines

- SVMs find a surface that separates "positive" and "negative" data points in high dimension. The "line" or surface is chosen such that it maximizes the "margin", i.e. the distance to the nearest of the data points. This "line" can be a polynomial separator but complexity should be penalized because of a tendency to over-fit the data.
- Radial Basis Function: put a positive or negative Gaussian at each training point and then sum.
- Each training point is a "vector", and some subset of these will form the "support" for the nonlinear separator – these are the "support vectors". Data points far from the separator have no influence on its location.
- If data cannot be perfectly separated, separator is positioned to optimize some combination of margin and penalty for misclassified points..
- In either case, calculation of the optimal separator is feasible – no problem getting stuck at local optima as with some other learning algorithms (e.g. EM).

Linear Discriminant Analysis

- Consider a case where we have two groups of data (in 2D), both are Gaussian with the same covariance, but different mean. The separator will be a line.

Quadratic Discriminant Analysis

- If instead the two Gaussians have different covariances, the optimal separator will be a quadratic function.

But if the data is not Gaussian, then neither of these methods will work. In particular, if the data is prone to outliers, they will heavily skew the mean and variance estimates. SVMs avoid the Gaussian assumption and are less susceptible to outliers.